Breaking Barriers in Physical-World Adversarial Examples: Improving Robustness and Transferability via Robust Feature

Yichen Wang[1,2,4,5,6], **Yuxuan Chou**[6], **Ziqi Zhou**[1,2,3,7], **Hangtao Zhang**[6], **Wei Wan**[1,2,4,5,6], **Shengshan Hu**[1,2,4,5,6], **Minghui Li**[8]

[1]National Engineering Research Center for Big Data Technology and System
[2]Services Computing Technology and System Lab
[3]Cluster and Grid Computing Lab
[4]Hubei Engineering Research Center on Big Data Security
[5]Hubei Key Laboratory of Distributed System Security
[6]School of Cyber Science and Engineering, Huazhong University of Science and Technology
[7]School of Computer Science and Technology, Huazhong University of Science and Technology
[8]School of Software Engineering, Huazhong University of Science and Technology
{wangyichen,yuxuanchou,zhouziqi, hangt_zhang,wanwei_0303, hushengshan, minghuili}@hust.edu.cn

Abstract

As deep neural networks (DNNs) are widely applied in the physical world, many researches are focusing on physical-world adversarial examples (PAEs), which introduce perturbations to inputs and cause the model's incorrect outputs. However, existing PAEs face two challenges: unsatisfactory attack performance (*i.e.*, poor transferability and insufficient robustness to environment conditions), and difficulty in balancing attack effectiveness with stealthiness, where better attack effectiveness often makes PAEs more perceptible.

In this paper, we explore a novel perturbation-based method to overcome the challenges. For the first challenge, we introduce a strategy Deceptive RF injection based on robust features (RFs) that are predictive, robust to perturbations, and consistent across different models. Specifically, it improves the transferability and robustness of PAEs by covering RFs of other classes onto the predictive features in clean images. For the second challenge, we introduce another strategy Adversarial Semantic Pattern Minimization, which removes most perturbations and retains only essential adversarial patterns in AEs. Based on the two strategies, we design our method Robust Feature Coverage Attack (RFCoA), comprising Robust Feature Disentanglement and Adversarial Feature Fusion. In the first stage, we extract target class RFs in feature space. In the second stage, we use attention-based feature fusion to overlay these RFs onto predictive features of clean images and remove unnecessary perturbations. Experiments show our method's superior transferability, robustness, and stealthiness compared to existing state-of-the-art methods. Additionally, our method's effectiveness can extend to Large Vision-Language Models (LVLMs), indicating its potential applicability to more complex tasks.

Code — https://github.com/CGCL-codes/RFCoA

1 Introduction

Deep neural networks (DNNs) have achieved significant milestones in various domains such as image recognition (Iandola et al. 2014), natural language processing (Achiam et al. 2023), and speech recognition (Maas et al. 2017). However, their inherent security issues have become increasingly prominent. One of the widely studied problems is the adversarial attack (Goodfellow, Shlens, and Szegedy 2015; Zhou et al. 2024a, 2023a; Song et al. 2025), where the adversary manipulates the model to output incorrect results by adding perturbations to the inputs. Previous works focus on adversarial examples (AEs) in the digital domain, which can be categorized into two approaches: perturbation-based methods (Madry et al. 2017; Zhou et al. 2024b), and patch-based methods (Brown et al. 2017; Casper et al. 2022). The former involve adding perturbations usually constrained by a specific norm, offering better stealthiness, while the latter involve applying elaborate patches to specific regions of the images, providing better attack performance but breaching the stealthiness.

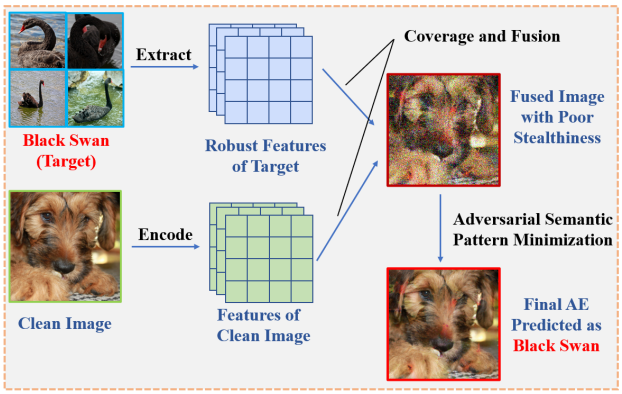

Figure 1: Our strategies Deceptive RF Injection and Adversarial Semantic Pattern Minimization.

With the application of DNNs in the physical world, such as autonomous driving (Zhang et al. 2024a) and facial recognition (Li et al. 2024), recent works pay more attention to AEs in real-world scenarios, known as physical-world adversarial examples (PAEs). Due to environmental factors in

the physical world (*e.g.*, distance, angles, and lighting), only a few perturbation-based methods (Jia et al. 2022; Ge et al. 2024) exhibit effective attacks in real-world scenarios, but they still suffer from poor transferability and lack robustness to changes of environmental factors. Current works on PAEs mainly focus on patch-based methods (Eykholt et al. 2018; Yang et al. 2020; Tan et al. 2023) and optical-based methods (Duan et al. 2021; Zhong et al. 2022; Wang et al. 2023a) that leverage beams or shadows in the physical world. However, both of them have inherent limitations (Wang et al. 2023b). Patch-based methods compromises the stealthiness to maintain attack performance, making PAEs more detectable. Optical-based methods are fragile to the variation of environmental factors, limiting its effectiveness to specific scenarios. In summary, existing PAEs face two challenges: the first is the unsatisfactory attack performance in real-world scenarios, *i.e.*, poor transferability and robustness, while the second is difficult trade-off between attack effectiveness and stealthiness.

Due to the inherent limitations of patch-based and optical-based methods, it is difficult to fundamentally address these challenges by them, so we explore a perturbation-based method to overcome the challenges. Recent works (Ilyas et al. 2019; Springer, Mitchell, and Kenyon 2021b) point out that existing perturbation-based methods concentrate merely on adversarially manipulating non-robust features (N-RFs), which are highly predictive (*i.e.*, playing a critical role in the model's prediction) but sensitive to perturbations and vary across models. Thus, the adversarial N-RFs fail to neither influence the model's prediction when faced with changes of environmental factors in real-world scenarios nor perceived by other black-box models (Wang et al. 2024a), which arises the first challenge. In contrast, there exists another type of predictive features known as robust features (RFs). They are strongly correlated with the image's semantics, robust to perturbations, and can be perceived by different models (Springer, Mitchell, and Kenyon 2021a; Benz, Zhang, and Kweon 2021). Therefore, we propose a novel strategy to overcome the first challenge, referred as Deceptive RF Injection, which involves covering RFs of other classes onto the predictive features in clean images. For the second challenge, we propose another strategy, Adversarial Semantic Pattern Minimization, which involves removing most perturbations and preserving only essential adversarial semantic patterns in AEs. Our strategies are illustrated in Fig. 1.

Based on above two strategies, we propose Robust Feature Coverage Attack (RFCoA) to generate PAEs with excellent transferability, robustness and stealthiness, which consists of Robust Feature Disentanglement and Adversarial Feature Fusion. In Robust Feature Disentanglement, we design an optimization process to extract RFs of the target class. During Adversarial Feature Fusion, we adopt the attention mechanism to fuse the RFs with clean images and optimize the attention weights of RFs to accurately covering predictive features in clean images, thus achieving targeted adversarial attack. Besides, according to the second strategy, we combine the minimal cognitive pattern approach (Huang et al. 2023b) to eliminate unnecessary perturbations and extract adversarial semantic patterns from fusion results by optimizing a pattern mask.

Method	Type	Setting	Trans.	Rob.	Steal.
AdvPatch	**Patch**	**White-box**	○	◐	○
TPA	**Patch**	**Black-box**	◐	●	○
AdvLB	**Optical**	**Black-box**	○	○	●
C/P Attack	**Patch**	**Black-box**	◐	●	○
ShadowAttack	**Optical**	**Black-box**	◐	○	●
RFLA	**Optical**	**Black-box**	◐	○	●
CleanSheet	**Perturbation**	**Black-box**	◐	◐	●
RFCoA (Ours)	**Perturbation**	**Black-box**	●	●	●

Table 1: Comparison among existing representative works on PAEs and our method. "●" indicates that the method performs well in the aspect and "◐" indicates the method shows some improvement but still remains mediocre.

We evaluate our method on ImageNet ILSVRC 2012 (Russakovsky et al. 2015) in both digtal and physical scenarios. The experimental results demonstrate that our method outperforms all existing state-of-the-art (SOTA) physical-world adversarial attacks in terms of transferability, robustness, and stealthiness. Furthermore, we also demonstrate the effectiveness of our method on large vision-language models (LVLMs) (Zhang et al. 2024b; Wang et al. 2024c), such as MiniGPT-4 (Zhu et al. 2023) and LLaVA (Liu et al. 2023), which indicates its potential in more complex scenarios and tasks.

In conclusion, the key contributions of our work are outlined as follows:

1) We provide a comprehensive review and summary of the challenges of existing PAEs and propose two novel strategies, Deceptive RF Injection and Adversarial Semantic Pattern Minimization, to fundamentally address the challenges.
2) Based on the proposed two strategies, we design a novel physical-world adversarial attack method RFCoA with high transferability, robustness and stealthiness, which consists of Robust Feature Disentanglement and Adversarial Feature Fusion.
3) Extensive experiments demonstrate the superiority of our method compared with existing SOTA methods. Additionally, we also demonstrate the effectiveness of our method on LVLMs, indicating its potential for applying to more complex scenarios.

2 Related Work

2.1 Adversarial Example in the Digital Domain

Adversarial examples (AEs) (Goodfellow, Shlens, and Szegedy 2015; Zhou et al. 2023b) are created by introducing imperceptible adversarial perturbations into images, leading to incorrect outputs of the model during the inference stage. Existing works on AEs in digital domain mainly focus on perturbation-based methods, such as FGSM (Goodfellow, Shlens, and Szegedy 2015), PGD (Madry et al. 2017), and C&W (Carlini and Wagner 2017). These methods effectively attack white-box models but face challenges in transferring

to black-box models. Furthermore, the introduced adversarial perturbations are fragile and tend to lose their effectiveness in real-world scenarios (Wang et al. 2023b; Zhou et al. 2025). Other patch-based methods, like AdvPatch (Brown et al. 2017) and DPatch (Liu et al. 2018), exhibit better robustness and have potential for application in the physical world, but the adversarial patches are too conspicuous, making them easy-to-detect.

2.2 Adversarial Example in the Physical World

More recently, an increasing number of works focus on deploying AEs in the physical world. Due to the fragility of perturbation-based methods to environmental factors, only a few works (*e.g.*, HA&NTA(Jia et al. 2022) and CleanSheet (Ge et al. 2024)) are effective in real-world scenarios, but their transferability and robustness remain poor. Some works (*e.g.*, RP2 (Eykholt et al. 2018), Copy/Paste Attack (Casper et al. 2022), T-Sea (Huang et al. 2023a), DOE (Tan et al. 2023), and etc.) involve carefully designing patches or camouflages and applying adaptive transformations, like Expectation Over Transformation (EOT) (Athalye et al. 2018), to enhance the robustness to physical-world perturbations. Despite achieving relatively good effectiveness on white-box models, they fail to remain satisfactory performance on black-box models. Moreover, they sacrifice the stealthiness of PAEs, making them more detectable, which is the inherent shortcoming of this kind of methods. Others like ShadowAttack (Zhong et al. 2022), AdvLB (Duan et al. 2021), and RFLA (Wang et al. 2023a) leverage optical perturbations like beam and shadows. While they exhibit excellent stealthiness, their inherent deficiency is the limited robustness, making them highly sensitive to the environments.

In summary, as shown in Tab. 1, existing works on PAEs face two challenges: unsatisfactory attack performance in real-world scenarios (*i.e.*, poor transferability and robustness), and difficulty in balancing attack effectiveness with stealthiness. Considering the inherent deficiencies of patch-based and optical-based methods, we design a perturbation-based method to overcome these challenges in this paper.

3 Methodology

3.1 Problem Definition

Notably, the attack we consider is in the black-box scenario, where the adversary can only access the dataset information, the output of the victim model f and cannot obtain the model's parameters or intermediate results. However, the adversary can employ several surrogate models with different structures from f trained on the dataset. This threat model is consistent with many existing works on PAEs (Tan et al. 2023; Huang et al. 2023a).

Considering that the perturbations from the physical world reduce the model accuracy, we choose to launch the targeted attack, which are more challenging than the untargeted attack. Given a classifier f, a clean image x and its label y, and a target class t, The adversary's goal is to create an adversarial example x' that satisfies Eq. (1):

$$\begin{aligned} min \quad & \|x' - x\| \\ s.t. \quad & f(x') = t \end{aligned} \tag{1}$$

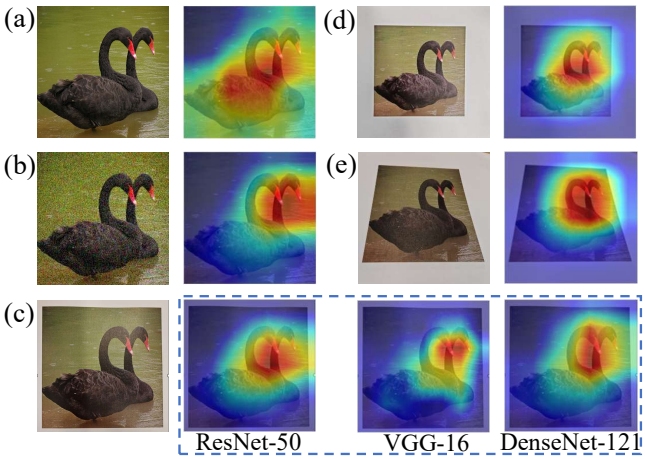

Figure 2: Attention maps calculated by Grad-CAM. (a) is the clean image in the digital world, (b) is the image added random noise in the digital world, and (c) to (e) are sampled in the physical world with various distances and angles. Notably, except for the special annotations, all models used to compute the attention maps are ResNet-50.

3.2 Intuition

Recent works (Ilyas et al. 2019; Wang et al. 2024b) point out that AEs generated by existing perturbation-based methods merely focus on adversarially manipulating non-robust features (N-RFs) that are highly predictive but brittle to perturbations and vary across different models, to cause the model's incorrect outputs. Due to environmental perturbations in real-world scenarios, the adversarial effectiveness of these N-RFs is significantly degraded, failing to influence the model's prediction. Furthermore, adversarial N-RFs are also difficult for other black-box models to perceive, let alone manipulate their outputs. Therefore, existing perturbation-based methods on PAEs exhibit deficiencies in both robustness and transferability.

(i) How to fundamentally improve the transferability and robustness of AEs? We need to leverage another type of predictive features that are more robust to perturbations and be perceived by multiple models, to launch attack. Fortunately, robust features (RFs) happen to satisfy both of these requirements perfectly (Springer, Mitchell, and Kenyon 2021a; Benz, Zhang, and Kweon 2021). Although these properties have been demonstrated in the digital domain, further exploration in the physical world is still necessary. Here, we sample images in both digital and physical worlds, input them into models, and utilize Grad-CAM (Selvaraju et al. 2017) to visualize the attention maps, as shown in Fig. 2. Comparing (a) and (b), in the image perturbed by noise, the model's attention shift from the feather texture of the black swan to the neck, head, and beak, indicating that RFs are distributed in these regions. According to (c) to (e), the model's attention maps largely overlap with those in (b), demonstrating RFs remain highly predictive and robust to changes of environmental conditions in the physical world. Additionally, in Fig. 2 (c), different models exhibit highly similar attention patterns for the physical-world sample, all

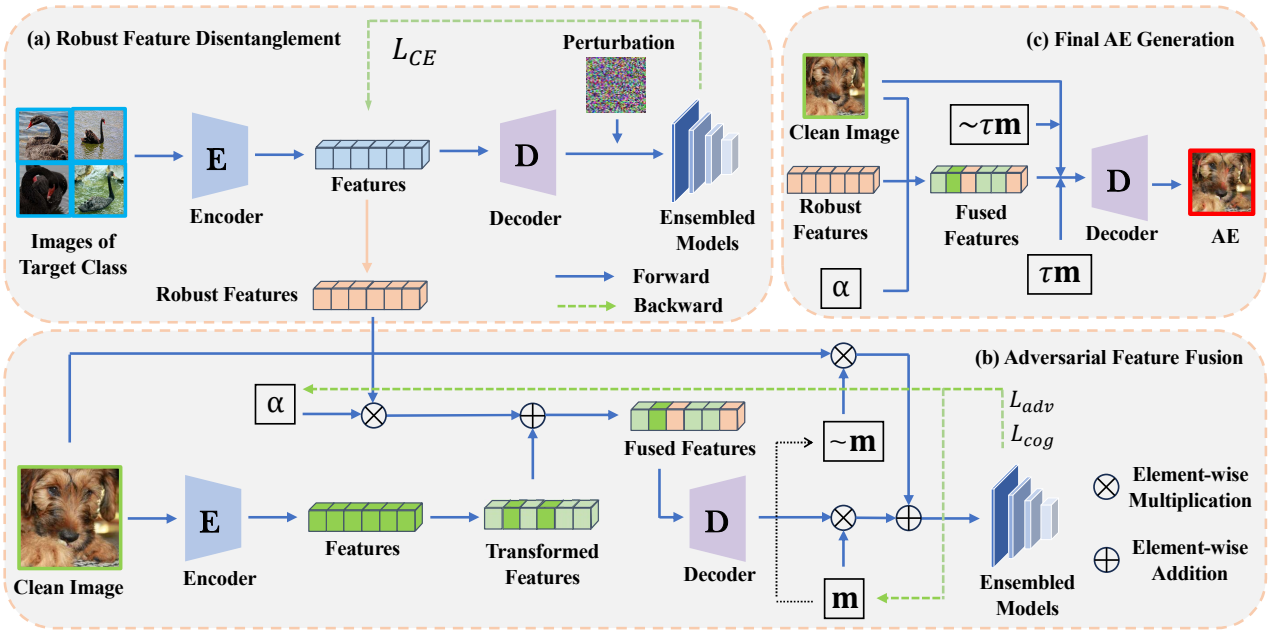

Figure 3: The overview of our method. (a) and (b) are the two modules of our method. After optimizing the α and $\mathbf{m}$ in (b), we calculate the final PAE by them through (c).

concentrating on RFs. The above observation indicates the properties of RFs also valid in the physical world. Thus, we can utilize them to fundamentally improve the transferability and robustness of AEs.

(ii) How to achieve better trade-off between the attack performance and stealthiness? Due to the perceptibility of RFs to humans, directly integrating them into clean images would compromise the stealthiness of AEs. Therefore, we dynamically adjust the weights of RFs during the fusion process and eliminate unnecessary perturbations to improve the stealthiness by minimizing adversarial patterns.

3.3 Overview of RFCoA

The overview of our method, RFCoA is shown in Fig. 3. It consists of two stages: Robust Feature Disentanglement and Adversarial Feature Fusion. In the first stage, we extract the RFs of the target class. In the second stage, we cover these RFs onto the clean image's predictive features. During this process, we optimize the weights α of the RFs in the fusion and the pattern mask $\mathbf{m}$ to minimize the visual difference between the AE and the clean image while maintaining attack performance. Finally, we execute the fusion process with the optimized α and $\mathbf{m}$ to generate the final AE, incorporating transparency τ to further improve stealthiness.

3.4 Robust Feature Disentanglement

Formally, the definition of robust features is as Eq. (2) :

$$\mathbb{E}_{(x,y)\in\mathcal{D}}\left[\inf_{\delta\in\Delta(x)} f\left(y\cdot E(x+\delta)\right)\right]\geq\gamma \tag{2}$$

where x and y are images and labels in the dataset $\mathcal{D}$, δ is the perturbation constrained in $\Delta(x)$, E is the robust feature extractor, $f(\cdot)$ is a function used to evaluate the correlation between features and labels, and γ is a predefined threshold.

The above definition implies that even affected by noise, the RFs in images remain highly predictive, allowing the model to rely on them for accurate predictions. In light of this, we design an optimization-based method to extract RFs. First, we employ the encoder of a pre-trained autoencoder to map images of the target class into the feature space to initialize the optimization target. Then, we reconstruct these features into images by the corresponding decoder, add noise, and input them into the model. After calculating the cross-entropy loss with the label, we iteratively optimize these features. Moreover, to enhance the universality of the extracted features, enabling them perceivable by models with different architecture, we ensemble several models and calculate the average loss.

Specifically, we formally express the optimization process for Robust Feature Disentanglement as Eq. (3):

$$\begin{aligned} f_t = \arg\min_{f} \quad & \frac{1}{N}\sum_{i=1}^{N} L_{CE}(\mathcal{M}_i(\mathbf{D}(f)+\delta), y_t) \\ \text{s.t.} \quad & f_0 = \mathbf{E}(x), \quad \|\delta\|_\infty \leq \epsilon \end{aligned} \tag{3}$$

where $\mathbf{E}$ and $\mathbf{D}$ are the encoder and decoder of the pre-trained autoencoder, x and y_t are images and label in the target class, L_{CE} represents the cross-entropy loss, N is the number of ensembled models, and $\mathcal{M}_i$ is the i-th model. Notably, f_0 is the initial value of the optimization target f, and the infinity norm of the perturbation δ is constrained by ϵ.

3.5 Adversarial Feature Fusion

After extracting RFs of the target class, we then fuse them into clean images. In this stage, we ensures the attack per-

formance in two aspects: weakening the predictive features in clean images, and overlaying RFs of the target class into original images. Due to the uneven distribution of predictive features in clean images, the weights of features at different positions should vary during fusion. Therefore, we adopt the attention mechanism (Vaswani et al. 2017). Initially, for clean images, we compute the spatial attention map in the feature space, which can reflect the distribution of predictive features. To simplify and clarify our method, we draw inspiration from the Grad-CAM, which employs the magnitude of gradients to assess the importance of features for prediction (Zhou et al. 2024b). Specifically, the process to obtain spatial attention maps can be expressed as Eq. (4) :

$$\mathcal{S} = F(|\nabla_{f_c} \frac{1}{N} \sum_{i=1}^{N} L(\mathcal{M}_i(\mathbf{D}(f_c), y))|) \tag{4}$$

where f_c is the representation of the clean image x_c in the feature space, *i.e.*, $\mathbf{E}(x_c)$, F is the sigmoid function that maps the absolute values of the gradients to range [0,1], and $\mathcal{S}$ is the calculated spatial attention map with the same shape as f_c. Given that positions with higher values in the attention map indicate concentrated predictive features, we transform the clean image's features as described in Eq. (5), thereby weakening the influence of predictive features of x_c in the subsequent fusion process.

$$f_c' = (1 - \mathcal{S}) \circ f_c \tag{5}$$

To accurately cover RFs of the target class onto predictive features of clean images distributed at different positions, we also employ the attention mechanism for the RFs in the fusion process, as shown in Eq. (6). x' is the fused image and α is the attention weights of RFs. Then we optimize α by minimizing the adversarial loss defined in Eq. (7).

$$x' = \mathbf{D}(\alpha \circ f_t + f_c') \tag{6}$$

$$L_{adv} = \frac{1}{N} \sum_{i=1}^{N} (w_1 L(\mathbf{M}_i(x'), y_t) - w_2 L(\mathbf{M}_i(x'), y_c)) \tag{7}$$

where w_1 and w_2 are pre-set weight parameters. The first term aims to make the fused image exhibit the semantics of f_t as much as possible, while the second term further weakening the influence of f_c.

However, the above fusion process does not consider the trade-off between the attack performance and stealthiness of the generated AEs. Here, we employ the minimal cognitive pattern (Huang et al. 2023b) to remove unnecessary corruption and preserve only essential adversarial patterns. The fusion process should be rewritten as:

$$x' = \mathbf{m} \circ \mathbf{D}(\alpha \circ f_t + f_c') + (1 - \mathbf{m}) \circ x_c \tag{8}$$

Notably, $\mathbf{m}$ is the pattern mask with the same shape as x_c and needs to be optimized along with α during the fusion process. Additionally, we introduce the cognitive loss L_{cog} to constrain the corruption to the clean image.

$$L_{cog} = w_3 \|\mathbf{m}\|_1 + w_4 TV(\mathbf{m}) - w_5 SSIM(x', x) \tag{9}$$

where $\|\cdot\|_1$ is the l_1 norm, $TV(\cdot)$ represents the total variation loss, $SSIM(\cdot)$ is the Structural Similarity Index Measure (SSIM) to measure the similarity between x' and x_c, and w_3 to w_5 are pre-set weights. The first item aims to eliminate unnecessary perturbations in the fusion results by minimizing the norm of pattern mask, while the latter two terms ensure the generated AEs have a smoother visual appearance and a higher similarity to the original image.

In conclusion, the whole optimization during the fusion can be expressed as Eq. (10) :

$$\arg\min_{\alpha, \mathbf{m}} \quad L_{adv} + L_{cog} \tag{10}$$

Furthermore, to facilitate direct control over the visual appearance of the AEs, we introduce a transparency parameter, τ, which serves as a weight factor on $\mathbf{m}$ when calculating the final result by Eq. (8).

4 Experiments

In this section, we evaluate our method in both digital and physical worlds and compare it with existing SOTA works on PAEs. In addition, we also discuss the attack performance on LVLMs and defenses strategies. The ablation studies and discussion of the attack performance under defenses are provided in the supplementary.

4.1 Setup

Dataset and classifiers. We conduct experiments on ImageNet ILSVRC 2012 (Russakovsky et al. 2015). In preprocessing, we resize and crop the images to 224x224. To comprehensively evaluate transferability, we select 12 commonly used models in image classification, including the ResNet (RN) series (He et al. 2016), Wide ResNet (WRN) series (Zagoruyko and Komodakis 2016), VGG series (Sengupta et al. 2019), DenseNet (DN) series (Iandola et al. 2014), GoogleNet (Szegedy et al. 2015), ShuffleNet (Zhang et al. 2018b), and Vision Transformer (ViT) (Vaswani et al. 2017). For each attack, we use three surrogate models: ResNet-50, VGG-16, and DenseNet-121, treating the others as black-box models.

Attack settings. We compare our method with existing SOTA physical-world adversarial attacks in image classification, including TPA (Yang et al. 2020), Copy/Paste Attack (C/P-A) (Casper et al. 2022), RFLA (Wang et al. 2023a) and CleanSheet (CS) (Ge et al. 2024). We utilize the official open-source code and select the settings claimed to achieve the best performance for evaluation. Details of our method's settings can be found in the supplementary.

Metrics. We adopt four metrics for evaluation: Clean Accuracy, Target Attack Success Rate (tASR) (Zhang et al. 2023), Structural Similarity Index Measure (SSIM) (Hore and Ziou 2010), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018a). Clean Accuracy measures the model's accuracy on clean inputs, while tASR represents the rate at which adversarial examples are misclassified as the target class. SSIM and LPIPS evaluate the stealthiness of AEs, with a smaller LPIPS value and a larger SSIM value suggesting better stealthiness.

Attacks \ Models	**RN-50***		RN-101		WRN-50		WRN-101		**VGG-16***		VGG-19	
	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.
Clean Acc.	0.76	0.60	0.77	0.62	0.79	0.64	0.79	0.68	0.72	0.56	0.72	0.61
C/P-A	0.88	0.37	0.12	0.07	0.27	0.20	0.14	0.08	0.78	0.49	0.62	0.37
TPA	0.88	0.41	0.10	0.05	0.12	0.11	0.07	0.05	0.95	0.55	0.65	0.39
RFLA	0.87	0.44	0.25	0.15	0.29	0.21	0.26	0.18	0.90	0.48	0.43	0.27
CS	0.84	0.46	0.51	0.22	0.33	0.19	0.18	0.11	**0.99**	0.51	0.89	0.35
Ours	**1.00**	**0.87**	**0.59**	**0.38**	**0.65**	**0.60**	**0.43**	**0.39**	**0.99**	**0.92**	**0.96**	**0.67**

Attacks \ Models	**DN-121***		DN-169		DN-201		ShuffleNet-v2		ViT-b32		GoogleNet	
	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.	Dig.	Phy.
Clean Acc.	0.74	0.58	0.76	0.63	0.77	0.65	0.76	0.66	0.76	0.63	0.70	0.59
C/P-A	0.69	0.32	0.17	0.13	0.20	0.14	0.17	0.12	0.04	0.02	0.16	0.10
TPA	0.93	0.40	0.15	0.10	0.06	0.04	0.07	0.06	0.08	0.05	0.03	0.03
RFLA	0.83	0.43	0.39	0.27	0.23	0.16	0.14	0.07	0.06	0.03	0.17	0.12
CS	0.98	0.42	0.71	0.36	0.68	0.31	0.17	0.11	0.04	0.02	0.11	0.08
Ours	**0.99**	**0.71**	**0.74**	**0.69**	**0.79**	**0.59**	**0.40**	**0.28**	**0.18**	**0.14**	**0.30**	**0.23**

Table 2: Quantitative results of attacks across various models in both digital and physical worlds. "Clean Acc." values denote the clean accuracy, while the other rows report tASR values. Notably, the models marked with "*" denote our surrogate models that are treated as white-box models, and the others are black-box models.

4.2 Attack Performance

Sampling. In the digital world, we randomly select 1000 images from ImageNet and set the attack target for each image to a randomly chosen class other than the original label. In the physical world, we randomly select 100 generated AEs, print them on 10cm x 10cm white paper, and photograph each with an iPhone 14 from a 10cm distance. The images are then resized to 224x224 for model inputs.The tASR results across different models are reported in Tab. 2.

As shown in Tab. 2, these attacks exhibit high tASR values on white-box models but vary significantly in black-box scenarios. Patch-based methods C/P Attack and TPA show poor transferability, with low tASR on most black-box models. RFLA performs relatively well on WRN series models but has limited transferability to other black-box models. CleanSheet shows improved transferability in the digital world, but suffers significant reduction on tASR values in physical-world scenarios. In contrast, our method achieves the highest tASR across all black-box models, including those with distinct architectures such as ViT-b32 and GoogleNet, demonstrating superior effectiveness and transferability in both digital and physical worlds.

4.3 Robustness and Stealthiness

Evaluation of robustness. Here, we consider two common perturbation factors in the physical world: the distance and the angle during sampling. To this end, we set three sampling distances of 10cm, 15cm, and 20cm, and three angles of 15°, 30°, and 45°. The results are presented in Tab. 4.

From Tab. 3, RFLA and CleanSheet are highly sensitive to variations in sampling distance and angles, with obvious degradation in tASR, which indicates that their robustness in the physical world is limited. In contrast, the patch-based attacks, C/P Attack and TPA, demonstrate better robustness in the physical world, with minimal changes in tASR. Our method exhibits the most pronounced robustness, maintaining a tASR of 0.22 on black-box models even under the most challenging 45°sampling condition, significantly surpassing the other methods.

Attacks \ Distances	10cm		15cm		20cm	
	W-b	B-b	W-b	B-b	W-b	B-b
C/P-A	0.41	0.14	0.36	0.13	0.30	0.09
TPA	0.45	0.10	0.41	0.09	0.36	0.07
RFLA	0.45	0.16	0.38	0.12	0.34	0.10
CS	0.46	0.19	0.43	0.16	0.37	0.12
Ours	**0.83**	**0.44**	**0.78**	**0.41**	**0.74**	**0.38**
Attacks \ Angles	15°		30°		45°	
	W-b	B-b	W-b	B-b	W-b	B-b
C/P-A	0.41	0.13	0.35	0.11	0.30	0.08
TPA	0.41	0.08	0.35	0.06	0.32	0.05
RFLA	0.41	0.13	0.27	0.06	0.19	0.03
CS	0.42	0.18	0.36	0.13	0.31	0.09
Ours	**0.79**	**0.39**	**0.73**	**0.35**	**0.56**	**0.22**

Table 3: Comparative results of average tASR under various distance and angles. "W-b" denotes white-box models, while "B-b" represents black-box models.

Methods	C/P-A	TPA	RFLA	CS	Ours
SSIM (↑)	0.56	0.68	0.84	0.77	**0.89**
LPIPS (↓)	0.59	0.36	0.20	0.25	**0.14**

Table 4: Comparative results of average SSIM and LPIPS values across different methods.

Evaluation of stealthiness. We select the AEs in Section 4.2 and compute their SSIM and LPIPS values relative to clean images in the digital domain to evaluate their stealthiness. The experimental results are presented in Tab. 4 and some physical-wrold visualization results are shown in Fig. 4.

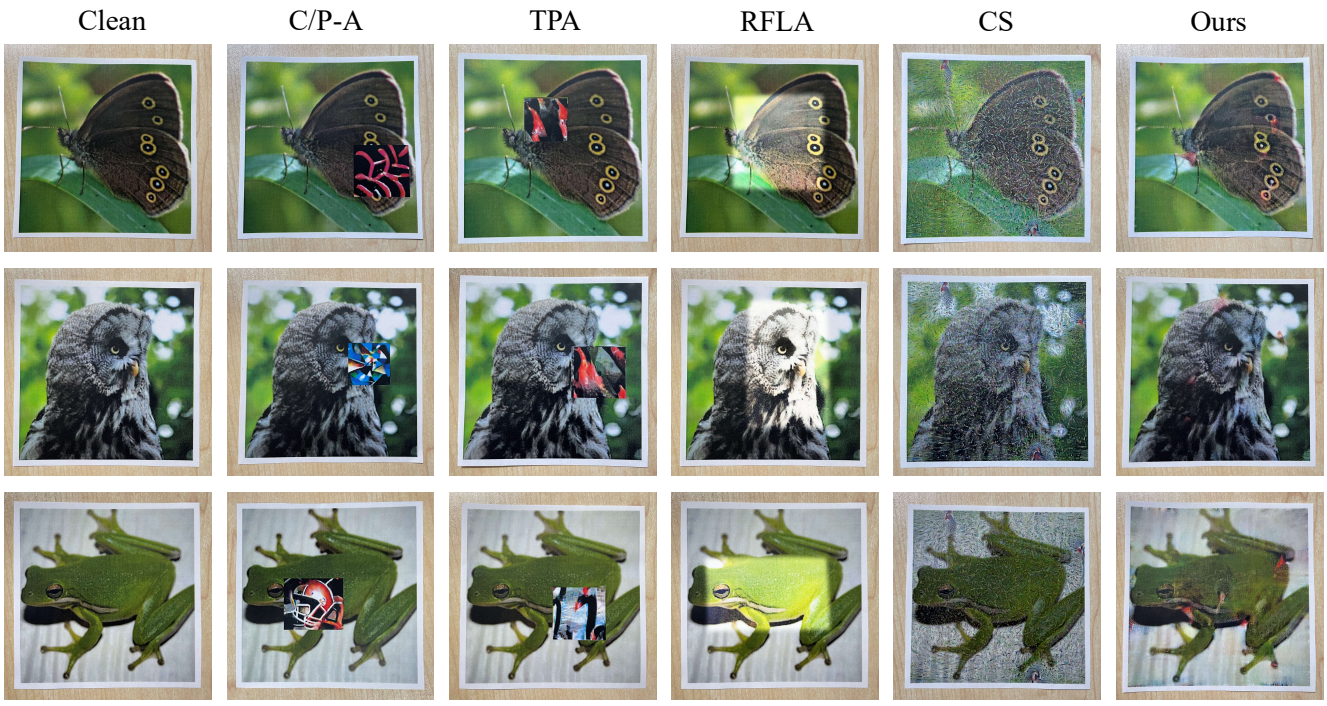

Figure 4: Visualization results of PAEs in the physical world.

For C/P Attack and TPA, due to the use of patches that significantly differ from the original images, AEs are easily perceived by humans, with SSIM values below 0.7. In contrast, RFLA, CleanSheet, and our method introduce mild perturbations, avoiding visually disruptive areas and thus achieving better stealthiness. Notably, our method also outperforms RFLA and CleanSheet in numerical results, demonstrating its superior stealthiness.

4.4 Attack against LVLMs

To explore the potential of our method on complex tasks and models, we apply AEs to LVLM-based Visual Question Answering (VQA) and image caption tasks.

Settings. We employ 100 AEs generated by our method to launch targeted attack against MiniGPT-4 (Zhu et al. 2023) and LLaVA (Liu et al. 2023) and record the average tASR values. Notably, the attack is considered successful if the LVLMs' response includes the target class or its synonyms. The prompts are set like " The image is from ImageNet. What is its class?" for VQA task and like " Please describe the image." for image description. Besides, we also enhance the prompts by appending the target class name or related terms at the end. The results are shown in Tab. 5. More visualization results and test screenshots are provided in the supplementary.

As shown in Tab. 5, our method outperforms in image caption tasks compared to VQA. This is likely because, in image caption, the model's attention is spread across the entire image, increasing the chances of recognizing adversarially injected RFs. Additionally, incorporating the target class name or related terms in textual prompts significantly boosts tASR. We hypothesize that these enhanced prompts guide the LVLM to focus more on the RFs of the target class, resulting in outputs more closely aligned with it. In conclusion, these results demonstrate the potential of our attack method to transfer to LVLM and multimodal tasks.

Prompts \ Models	MiniGPT-4		LLaVA	
	VQA	Caption	VQA	Caption
Normal	0.36	0.44	0.32	0.41
Enhanced	0.57	0.65	0.50	0.57

Table 5: Quantitative results of our method on LVLMs.

5 Conclusion

In this work, we propose two novel strageties Deceptive RF Injection and Adversarial Semantic Pattern Minimization to fundamentally overcome the challenges of existing PAEs. Based on the strategies, we design a perturbation-based attack RFCoA to craft PAEs with excellent transferability, robustness, and stealthiness. Experimental results demonstrate the superiority of our method compared with existing SOTA works in both digital and physical scenarios. Moreover, our method shows effectiveness on LVLMs, which indicates the potential of our attack to transfer to more complex tasks.

Acknowledgments

Minghui's work is supported by the National Natural Science Foundation of China (Grant No. 62202186). Shengshan's work is supported by the National Natural Science Foundation of China (Grant No. 62372196). Wei Wan is the corresponding author.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing Robust Adversarial Examples. In *Proceedings of the 35th International Conference on Machine Learning, (ICML'18)*, 284–293.

Benz, P.; Zhang, C.; and Kweon, I. S. 2021. Batch Normalization Increases Adversarial Vulnerability and Decreases Adversarial Transferability: A Non-Robust Feature Perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*, 7818–7827.

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Carlini, N.; and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *Proceedings of 2017 IEEE Symposium on Security and Privacy (S&P'17)*, 39–57.

Casper, S.; Nadeau, M.; Hadfield-Menell, D.; and Kreiman, G. 2022. Robust Feature-Level Adversaries are Interpretability Tools. In *Proceedings of the 35th Advances in Neural Information Processing Systems (NeurIPS'22)*, 33093–33106.

Duan, R.; Mao, X.; Qin, A. K.; Chen, Y.; Ye, S.; He, Y.; and Yang, Y. 2021. Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 16062–16071.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, 1625–1634.

Ge, Y.; Wang, Q.; Huang, H.; Li, Q.; Wang, C.; Shen, C.; Zhao, L.; Jiang, P.; Fang, Z.; and Zhang, S. 2024. Hijacking Attacks against Neural Networks by Analyzing Training Data. *arXiv preprint arXiv:2401.09740*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 770–778.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR'10)*, 2366–2369.

Huang, H.; Chen, Z.; Chen, H.; Wang, Y.; and Zhang, K. 2023a. T-SEA: Transfer-Based Self-Ensemble Attack on Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, 20514–20523.

Huang, H.; Ma, X.; Erfani, S. M.; and Bailey, J. 2023b. Distilling Cognitive Backdoor Patterns within an Image. In *Proceedings of the 11th International Conference on Learning Representations (ICLR'23)*.

Iandola, F.; Moskewicz, M.; Karayev, S.; Girshick, R.; Darrell, T.; and Keutzer, K. 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *Proceedings of the 32nd Advances in Neural Information Processing (NeurIPS'19)*, 125–136.

Jia, W.; Lu, Z.; Zhang, H.; Liu, Z.; Wang, J.; and Qu, G. 2022. Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems. In *Proceedings of the 29th Annual Network and Distributed System Security Symposium (NDSS'22)*.

Li, M.; Wang, J.; Zhang, H.; Zhou, Z.; Hu, S.; and Pei, X. 2024. Transferable Adversarial Facial Images for Privacy Protection. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24)*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *Proceedings of the 36th Advances in Neural Information Processing Systems (NeurIPS'23)*.

Liu, X.; Yang, H.; Liu, Z.; Song, L.; Li, H.; and Chen, Y. 2018. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*.

Maas, A. L.; Qi, P.; Xie, Z.; Hannun, A. Y.; Lengerich, C. T.; Jurafsky, D.; and Ng, A. Y. 2017. Building DNN acoustic models for large vocabulary speech recognition. *Computer Speech & Language*, 41: 195–213.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; and Bernstein, M. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'17*, 618–626.

Sengupta, A.; Ye, Y.; Wang, R.; Liu, C.; and Roy, K. 2019. Going deeper in spiking neural networks: VGG and residual architectures. *Frontiers in neuroscience*, 13: 95.

Song, Y.; Zhou, Z.; Li, M.; Wang, X.; Deng, M.; Wan, W.; Hu, S.; and Zhang, L. Y. 2025. PB-UAP: Hybrid Universal Adversarial Attack For Image Segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'25)*.

Springer, J. M.; Mitchell, M.; and Kenyon, G. T. 2021a. Adversarial Perturbations Are Not So Weird: Entanglement of Robust and Non-Robust Features in Neural Network Classifiers. *CoRR*, abs/2102.05110.

Springer, J. M.; Mitchell, M.; and Kenyon, G. T. 2021b. A Little Robustness Goes a Long Way: Leveraging Robust Features for Targeted Transfer Attacks. In *Proceedings of the 34th Advances in Neural Information Processing Systems (NeurIPS'21)*, 9759–9773.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going Deeper With Convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'15)*.

Tan, W.; Li, Y.; Zhao, C.; Liu, Z.; and Pan, Q. 2023. DOEPatch: Dynamically Optimized Ensemble Model for Adversarial Patches Generation. *arXiv preprint arXiv:2312.16907*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 30th Advances in Neural Information Processing Systems (NerIPS'17)*.

Wang, D.; Yao, W.; Jiang, T.; Li, C.; and Chen, X. 2023a. RFLA: A Stealthy Reflected Light Adversarial Attack in the Physical World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, 4455–4465.

Wang, H.; Deng, Y.; Yoo, S.; and Lin, Y. 2024a. Exploring robust features for improving adversarial robustness. *IEEE Transactions on Cybernetics*.

Wang, J.; Wang, D.; Hu, J.; Wu, S.; Jiang, T.; Yao, W.; Liu, A.; and Liu, X. 2023b. Adversarial Examples in the Physical World: A Survey. *CoRR*, abs/2311.01473.

Wang, X.; Li, M.; Liu, W.; Zhang, H.; Hu, S.; Zhang, Y.; Zhou, Z.; and Jin, H. 2024b. Unlearnable 3D Point Clouds: Class-wise Transformation Is All You Need. In *The 38th Conference on Neural Information Processing Systems (NeurIPS'24)*.

Wang, X.; Pan, H.; Zhang, H.; Li, M.; Hu, S.; Zhou, Z.; Xue, L.; Guo, P.; Wang, Y.; Wan, W.; et al. 2024c. TrojanRobot: Backdoor Attacks Against Robotic Manipulation in the Physical World. *arXiv preprint arXiv:2411.11683*.

Yang, C.; Kortylewski, A.; Xie, C.; Cao, Y.; and Yuille, A. 2020. Patchattack: A black-box texture-based attack with reinforcement learning. In *Proceedings of the 16th European Conference on Computer Vision (ECCV'22)*, 681–698.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, H.; Hu, S.; Wang, Y.; Zhang, L. Y.; Zhou, Z.; Wang, X.; Zhang, Y.; and Chen, C. 2024a. Detector collapse: Backdooring object detection to catastrophic overload or blindness. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence, IJCAI'24*.

Zhang, H.; Yao, Z.; Zhang, L. Y.; Hu, S.; Chen, C.; Liew, A.; and Li, Z. 2023. Denial-of-Service or Fine-Grained Control: Towards Flexible Model Poisoning Attacks on Federated Learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI'23*.

Zhang, H.; Zhu, C.; Wang, X.; Zhou, Z.; Yin, C.; Li, M.; Xue, L.; Wang, Y.; Hu, S.; Liu, A.; et al. 2024b. BadRobot: Manipulating Embodied LLMs in the Physical World. *arXiv preprint arXiv:2407.20242*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018b. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18)*, 6848–6856.

Zhong, Y.; Liu, X.; Zhai, D.; Jiang, J.; and Ji, X. 2022. Shadows Can Be Dangerous: Stealthy and Effective Physical-World Adversarial Attack by Natural Phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22)*, 15345–15354.

Zhou, Z.; Hu, S.; Li, M.; Zhang, H.; Zhang, Y.; and Jin, H. 2023a. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'23)*, 6311–6320.

Zhou, Z.; Hu, S.; Zhao, R.; Wang, Q.; Zhang, L. Y.; Hou, J.; and Jin, H. 2023b. Downstream-agnostic adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, 4345–4355.

Zhou, Z.; Li, B.; Song, Y.; Hu, S.; Wan, W.; Zhang, L. Y.; Yao, D.; and Jin, H. 2025. NumbOD: A Spatial-Frequency Fusion Attack Against Object Detectors. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI'25)*.

Zhou, Z.; Li, M.; Liu, W.; Hu, S.; Zhang, Y.; Wan, W.; Xue, L.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024a. Securely Fine-tuning Pre-trained Encoders Against Adversarial Examples. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP'24)*.

Zhou, Z.; Song, Y.; Li, M.; Hu, S.; Wang, X.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024b. Darksam: Fooling segment anything model to segment nothing. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS'24)*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.